DOCUMENT RESUME

ED 233 048 TM 830 479

AUTHOR Capie, William
TITLE Methodological Issues in Combining Teacher Performance Data for Decision-Making.
PUB DATE 13 Apr 83
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Check Lists; Decision Making; Educational Assessment; *Evaluation Methods; *Measurement Techniques; Rating Scales; *Standards; State Programs; *Teacher Certification; *Teacher Evaluation; *Teacher Qualifications; Validity

ABSTRACT

The discussed and illustrated principles regarding teacher performance data and decision making are common sense ideas that have been developed based on a good deal of thought about the definition of minimally acceptable job performances. The presented scoring systems are ways of describing how much a teacher must do in order to be scored positively, i.e., certified or continued in employment. The purpose of the assessment should be considered when the decision-making structure is planned. Since the pattern and nature of the requirements will certainly influence the demands that will be put on the teachers, developers of performance assessment systems would do well to consider how many criteria, categories, and sub-categories a teacher can fail to demonstrate, and still be called acceptable. Generalizability procedures can help determine the effect of alternative procedures, such as varying number of raters, occasions, or items. Whole scores may be more defensible for certification but inappropriate for staff development. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED233048

Methodological Issues in Combining Teacher Performance Data for Decision-Making

William Capie

Department of Science Education

University of Georgia

Athens, Georgia 30602

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. Capie

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 830 425

A paper presented at the Annual Meeting of the American Educational Research Association, Montreal, April 13, 1983.

Methodological Issues in Combining Teacher

Performance Data for Decision-Making

Teacher evaluation is "in!" This observation certainly is not news to anyone who follows the development of educational policy at state and local levels. While teacher evaluations have long been a part of personnel activities in most school districts, the quality and quantity of assessment has changed immensely in the last decade. Many of us grew up professionally with a ritual of an annual observation by a principal, or perhaps with an "evaluation" that was based on no observation at all. Typically these evaluations consisted of simple checklists, hastily constructed rating scales, or a short narrative which had little to do with any personnel action that might have been necessary. The changes of the seventies and eighties are most significant because the evaluation data that are being generated are being used to make important decisions.

The most visible uses of assessment data may be the statewide certification programs which seem to be proliferating, particularly in the southeast. Statewide assessment activities in Georgia, South Carolina, Florida, Kentucky, Oklahoma and Virginia all have elements of admission to the teaching profession and providing staff development. The Georgia program, which is the most mature, is almost exclusively a certification program although the data may be useful for staff development purposes. Although the South Carolina legislation states that data are to be used for employment decisions, their process appears to have many of the same purposes as Georgia's. The decision-making structure in Oklahoma and Florida programs is not well-described and the Virginia program is in its embryonic stages. The uses of teacher evaluation data at the school and district level are, or should

be, quite different than data collected as part of a statewide program.

The uses of the assessment data will be of great importance in determining what is to be measured and how the measurements will be made. The "what is to be measured?" issue is validity, of course. Answers in the statewide programs have been somewhat variable. Some, like Georgia and South Carolina, measure teaching dimensions that the state's educators have endorsed as important skill areas for teachers. Some, like Georgia and Florida, have attempted to document the effects of the various teaching dimensions in the professional literature. Georgia has even related its assessment measures to several different pupil outcome variables. At the very least, all of the statewide assessment schemes for beginning teachers seem to possess a "common sense" face validity and some have a much more systematic base of support in the research literature and criterion-related outcome studies.

"The how will measurements be made?" issue has not been received nearly as much attention in education related fields as it has in other areas. The literature of personnel psychology is replete with studies of various measurement methodologies applied to the performance assessment of personnel. Common methodologies include ranking, pair comparison, trait ratings, graphic rating scales, checklists, mixed standard scales, and behaviorally anchored rating scales (Jacobs, 1982). Other formats that are sometimes used include forced choice, essay narratives and critical incidents. Of these possibilities only checklists, mixed standard rating scales and behaviorally anchored rating scales are appropriate for the certification statewide

programs and dismissal decisions. The others suffer from lack of validity or the fixed standard that may be necessary in making personnel decisions.

Rosenshine and Furst in The Second Handbook on Research on Teaching (1973) describe a number of methodologies for measuring the classroom performance of teachers. Included in their list are the expressions "category systems" and "sign systems" which raise important questions about the frequency of performance of certain behaviors and the inference level required of observers. Although they point out that rating systems "are no longer limited to high inference items" (p. 133), there remains a persistent notion that sign systems or category systems are more objective and more suitable for teacher assessment because they require less inference by observers. However, data that are collected in typical sign of category systems with event or time sampling are difficult to apply in performance assessments because of uncertainties in scoring and setting meaningful criteria. Interestingly, to me anyhow, when Rosenshine and Furst described the purposes of observational systems in the classroom, they did not include the assessment of personnel! The focus of this symposium illustrates a considerable change in the ten years since their work appeared.

Conventional wisdom suggests that Behaviorally Anchored Rating Scales (BARS) are the most effective assessment format for reducing errors such as leniency and halo and for enhancing reliability in assessing personnel performance. Rating scales, which may be viewed as "a retrospective synthesis by one individual of the efforts or performance of another" (Landy & Farr, 1980, p. 74), were used for many years before the notion of behavioral anchors was introduced

in the early nineteen-sixties. Jacobs, Kafry and Zedeck (1980, p. 602) cite the strength of BARS as the definition of levels of performance. Landy and Farr recognize this goal in BARS but cite the continuing problem of defining mid-point anchors (p. 83). The intuitive appeal for the BARS for certification applications is strong in that clear performance dimensions can be constructed and essential levels of performance can be identified.

Conventional wisdom favoring BARS has been challenged by recent research involving various measurement methodologies. (Jacobs, 1982; Borman, 1982) Borman points out that

> This does not mean that performance rating scales can be developed carelessly. Concensus in the performance appraisal literature is that small variations in the rating format are not important as long as the scales are developed carefully. Thus, careful preparation of performance rating scales is probably critical for generating high-quality, relatively accurate, and error-free performance evaluations.

Jacobs, et al add to these concerns the need to keep the purposes of the assessment at the forefront as planning and development occur.

Careful planning of the intended use of the instrument also is an important area where results might be improved. Cooper (1981) cites a number of possible procedures which might reduce halo errors. He reinforced the ideas of Jacobs and Borman by indicating that improved scales are an important priority. Training can also improve the use of scales. Other suggestions which are reflected in the use of the instrument include providing more data (i.e. more observation time to observers), using multiple observers, assessing irrelevant categories first to help purge confusing data from the observers decision-making process, and assessing from current exposure (i.e. assessing the most recent observation of the individual and not a global or overall

assessment). Several of these proposals imply that increased amounts of data will be gathered and incorporated into the decision-making process.

The implications of these recommendations are enormous. Unfortunately, the literature contains few guidelines for aggregating data. Whenever there are possibilities of multiple observers, multiple occasions, various aggregations of items and sub-items, and the like, the number of variations in scoring procedures and decision-making is large. Graphic or other numerical scales lend themselves to summing, hence the behaviorally summated scales (BSS). Capie (1982) has shown that highly similar results are when scale scores are summed and when they are reduced to a dichotomous decision and then summed. Jacobs (1982) notes that checklists, unlike rating scales, produce a "derived score" which he claims can surprise observers when the individual being assessed shows up differently in the derived scores than the observers expected in their holistic or intuitive judgements. Such a phenomenon may not be undesirable if the system is sound and diminishes rater bias. Nevertheless, any system which involves the aggregation of assessment data will yield a derived score. And, the system for deriving the score(s) should be consistent with the purpose of the assessment and at least as well-conceived as the instrument. Drawing scores from scales and subscales is a complex process which raises many issues about the meaning of and weight of the data.

Many of the issues that I am about to raise can be applied equally to checklists or to BARS. Therefore, the assessment methodology is not particularly important so long as it meets the needs of the local situation. For example, in Georgia both BARS and checklists are used within the same instrument. Both types of items have specified minimum

levels, since teachers must reach a specified point on the rating scale or demonstrate a specified number of items on the checklist before receiving credit for an item. The balance of the paper is a review of issues that must be resolved as performance data are aggregated for decision-making.

Number of Decision-Making Units

The number of decision-making units to be included in an assessment process is a multi-level issue. The initial question revolves around the holistic purpose of the assessment. The concern might be as broad as "Can the individual demonstrate an adequate level of teaching skill?" If this broad question is appropriate, then a single yes/no decision may be all that is necessary. A teacher could pass or fail the whole test. Alternatively, there may be a need to deal with more detailed decisions such as "Can the person demonstrate each of these three, or nine, or fourteen, or thirty-three, or even four thousand teaching skills?" In Georgia we call these our "fourteen essential competencies." South Carolina has dubbed similar units "five Performance Dimensions." Dade County has seven "Categories!" Examples of the kinds of things that might be included at this level are "Plans Instruction," "Administers Class Routines," and "Maintains a Positive Learning Environment." In general, these major decision categories should be required of all candidates who are being assessed. They might be demonstrated on a single occasion or perhaps on different occasions depending on the purpose of the assessment. A relatively small number of these elements should be included in the assessment, certainly less than four thousand. Each of these divisions will have to be defined by sub-behaviors, a practical concern as the system is developed. The number of divisions at this level will have implications for the

number of repeat assessment and carry-overs, a practical concern when the program is operational.

Typically each of the major categories in an assessment system is partitioned into sub-categories. Thus, "Plans Instruction" may be defined by additional criteria, and so on. I do not know of a system with more than four levels of specificity. A perfectly symmetrical system where each major category is defined by the same number of sub-categories, each defined by the same number of sub-sub-categories, is convenient. But, such an arrangement is not necessary. I am going to use a symmetrical arrangement to illustrate some general principles, however. This combination happens to come from the Teacher Assessment and Development System which we have used in Dade County, Flordia. The content outline is included in Figure 1.

There is a total of thirty-two teaching behavior decisions that are included. Each represents a dichotomous decision that is made by an observer. (Each is indicated by a lower case letter). They are partitioned into eight units called sub-indicators. (Each is indicated by a numeral). These eight are divided into four indicators, the largest division within the category. (Each is indicated by a capital letter.) These thirty-two items are not a random set since there is some logic and research evidence to support the content of each of these various divisions. For the sake of convenience I have chosen to work with an arbitrary standard of "three-fourths of the sub-decisions must be positive before a supraordinate category will be scored positively." The tension that exists as standards are set is between the desire to be fair to the teacher by allowing some room to slip and some room for observer error and the desire to have the teacher demonstrate all of the numerous important components of effective teaching. Obviously,

Methodological Issues in Combining Teacher Performance Data for Decision-Making

William Capie
212 Aderhold Hall
University of Georgia
Athens, Georgia 30602

	Sampling Teaching Behaviors	Sampling Indicators (any three)	Sampling Sub-indicators (any six)	Sampling Teaching Behaviors within Indicators	Sampling Teaching Behaviors within Sub-indicators
A 1a	Any 24 out of 32	Any 6 of 8	Any 3 of 4	Any 6 of 8	Any 3 of 4
b					
c					
d					
2a			Any 3 of 4		Any 3 of 4
b					
c					
d					
B 1a		Any 6 of 8	Any 3 of 4	Any 6 of 8	Any 3 of 4
b					
c					
d					
2a			Any 3 of 4		Any 3 of 4
b					
c					
d					
C 1a		Any 6 of 8	Any 3 of 4	Any 6 of 8	Any 3 of 4
b					
c					
d					
2a			Any 3 of 4		Any 3 of 4
b					
c					
d					
D 1a				Any 6 of 8	Any 3 of 4
b					
c					
d					
2a					Any 3 of 4
b					
c					
d					

Figure 1. Possible scoring systems for one data set.

this decision is an important one that should be made with a great deal of care as assessment systems are designed. Possibilities for making category decisions by applying the three-fourths criterion to this data matrix include the following:

1. Sampling teaching behaviors where any twenty-four of the thirty-two teaching behaviors may be demonstrated.

This is a rather lenient criterion in that there are few restrictions as to how an acceptable performance will be demonstrated. In fact, a teacher might be quite weak in certain areas and compensate for the weakness by showing up as quite strong in other areas.

2. Sampling indicators where six of the eight teaching behaviors in any three of the four indicators may be required.

This requirement is even more relaxed than the first since it can be met with as few as eighteen of the teaching behaviors observed or credited.

3. Sampling sub-indicators where three of the four teaching behaviors in any six of the eight sub-indicators may be required.

This requirement is also more relaxed than the first and is not substantially different than the second in terms of rigor.

4. Sampling teaching behaviors within sub-indicators where six of the eight teaching behaviors in all four indicators may be required.

This criterion is more rigorous than any one of the other three. However, it allows the teacher the possibility of failing to demonstrate two behaviors in either sub-indicator in each of the indicators. Thus, it allows a strong performance in one sub-indicator

to compensate for a weak performance on another.

5. Sampling teaching behavior with sub-indicators where three of the four teaching behaviors may be required in all eight of the sub-indicators.

This criterion is the most rigorous of the set since the teacher has to meet a minimum level in each of the eight sub-indicators. It does not allow the possibility of a strong performance in one sub-indicator to compensate for a weak performance on the other.

One other consideration that may be desirable in designing a system such as this one is the importance of certain teaching behaviors. The ideal arrangement is to have four equally important teaching behaviors defining a sub-indicator so that the set can be sampled. Then, three out of four might be considered an adequate sample in order to give credit for the sub-indicator. However, sometimes one teaching behavior seems more critical than the others in the set. In such cases, the decision rule might be modified so that the "essential" teaching behaviors must be demonstrated before giving credit for a supraordinate category. Let's assume that the first teaching behavior in each sub-indicator must be observed before a positive decision can be made about the next supraordinate category and then see how the earlier decision rules are influenced.

1. When teaching behaviors are sampled, the eight teaching behaviors specified as essential and sixteen of the remainder must be demonstrated in order to have twenty-four of the thirty-two.

The effect of the essential teaching behaviors in this case is great. If the teacher fails to demonstrate any one or if an observer fails to recognize and check any one, then the teacher will be unacceptable in

the entire category.

2. When indicators are sampled the teacher must demonstrate both of the essential teaching behaviors and four others in three of the four indicators.

Thus, this sytem provides some room for slippage by the teacher. Although there is room for the teacher to fail to demonstrate two essential teaching behaviors and still receive credit for the category, both of the deficiencies must occur in the same indicator.

3. When sub-indicators are sampled the teacher must demonstrate the essential teaching behavior and two others in six of the eight sub-indicators.

This system is somewhat more flexible than the previous example in that a teacher can fail to demonstrate any two of the essential teaching behaviors and still be considered acceptable overall.

4. When teaching behaviors are sampled within sub-indicators, the two essential teaching behaviors and four others will be required in each of the four indicators if the teacher is to be acceptable in the category.

Like the first system, this one will result in a negative decision if any one of the essential teaching behaviors is unchecked.

5. When teaching behaviors are sampled within sub-indicators, the essential teaching behavior and two others will have to be observed in each sub-indicator if the teacher is to be acceptable in the category.

Like the first and fourth systems, this one will yield an unacceptable decision if any one of the essential teaching behaviors is not credited.

In summary, the pattern and nature of the requirements certainly influence the demands that will be put on teachers and the "clout" of the requirements. In these sorts of systems where all of the behaviors are considered desirable the sampling process must be examined carefully. Initially, when there was no concept of essential teaching behaviors, all five of the systems provided some room for slippage on the part of teachers and observers. Some may argue that first three examples allowed so much room that the integrity of the indicators and sub-indicators might have been lost. When the notion of essential teaching behaviors was imposed on the set of systems, these last two alternatives became quite rigid and perhaps unfair. While allowing a subset of "essential items" seems contradictory, doing so may be necessary to allow for rating errors and erratic performances. Clearly, if any one teaching behavior or very basic decision unit is a necessary condition for an acceptable decision in an entire category, that situation should be examined carefully lest it have undesirable effects.

Combining Multiple Data Sets

Few decisions about the adequacy of teaching performances should be based on one set of data, particularly when there are likely to be adverse effects of an unacceptable or negative decision. Consequently, many assessment models incorporate multiple observers and/or multiple assessment occasions. The simplest possibility is having two sets of data whether they are gathered by two observers at the same lesson or the same observer after two different lessons. The Georgia program provides for three or six sets of data to be combined in a single decision for each competency. In many respects the data aggregation problems are similar to the simple cases that were presented earlier.

Basically, there are three types of combination procedures that might be used: replication of single assessment strategies, pooling within sub-categories, or sampling sub-categories. These are summarized in Figure 2.

Replication of Single Assessment Procedures

Any one of the five procedures that were outlined earlier could be applied to both data sets independently. Then the teacher might be required to meet the criterion on both occasions or in the eyes of both observers. In general, the strengths and weaknesses of each of the five procedures is the same for multiple data sets as they are for the single data sets. Thus, demonstrating twenty-four of the thirty-two teaching behaviors on each of two occasions is the most lenient of the procedures. Of the three types of combination procedures replication is the most demanding, however. It allows one weak lesson or one harsh observer to penalize the teacher when an over-all decision is made.

Pooling within Sub-categories

Combining data within sub-categories can have many advantages in terms of fairness to the teacher while holding up standards to reasonable levels. By this procedure, we look at all of the teaching behaviors that were scored across the two occasions (or the two observers). Thus, we could pool within sub-indicators, making a total of eight teaching behaviors in each. Or, we could pool within indicators, creating a total of sixteen teaching behaviors in each. The combination criteria would then be six of the eight behaviors needed to make a positive decision for a sub-indicator or twelve of sixteen necessary to make a positive decision for an indicator. The difference in the two procedures is keeping or losing the identity of the sub-indicators. Once the data have been combined in this way, the category decision could be one of

	Pooling				Sampling			
	In Sub-indicators		In Indicators		Teaching Behaviors		Sub-indicators		Indicators	
	*Observer One	Observer Two	Observer One	Observer Two	Observer One	Observer Two	Observer One	Observer Two	Observer One	Observer Two
A1a	Any 6 of 8		Any 12 of 16		Any 48 of 64		Any 3 of 4	Any 3 of 4	Any 6 of 8	Any 6 of 8
b										
c										
d										
2a	Any 6 of 8						Any 3 of 4	Any 3 of 4		
b										
c										
d										
B1a	Any 6 of 8		Any 12 of 16				Any 3 of 4	Any 3 of 4	Any 6 of 8	Any 6 of 8
b										
c										
d										
2a	Any 5 of 8						Any 3 of 4	Any 3 of 4		
b										
c										
d										
C1a	Any 6 of 8		Any 12 of 16				Any 3 of 4	Any 3 of 4	Any 6 of 8	Any 6 of 8
b										
c										
d										
2a	Any 6 of 8						Any 3 of 4	Any 3 of 4		
b										
c										
d										
D1a										
b										
c										
d										
2a										
b										
c										
d										

*This could be Occasion one and Occasion two.

Figure 2. Alternative Procedures for Combining Two Data Sets.

the alternatives that were presented earlier--all sub-indicators or all indicators, six sub-indicators or three indicators, whichever seemed appropriate. Pooling data tends to have the benefit of preserving the identity of the sub-units, either indicator or sub-indicator, while providing "room" for the teacher. There is room for slightly uneven performances so that strong performances can compensate for weak ones, and also room for observer error without undue penalty.

Sampling Sub-categories

The sample data set that we have been using has three levels of sub-category--indicators, sub-indicators and teaching behaviors. Any of these might be sampled as the data are aggregated across two observations or two raters. The simplest case is sampling at the teaching behavior level, where the criterion would become forty-eight of the sixty-four teaching behaviors must have been scored positively in order to earn credit for the category. This is clearly a generous procedure for aggregating data since one very weak lesson could easily be compensated by a good lesson. Similarly, the effects of a "slightly biased observer" are most notable with this procedure.

Decisions about sub-indicators or indicators from each observation or each rater can also be incorporated into a reasonable process. Thus, the sixteen sub-indicator decisions can be examined to determine if twelve of them are positive. Or, we could examine the indicators to see if six of them had been scored positively. Sampling has the benefit of retaining the identity of the sub-categories while allowing some flexibility for the teacher. It is somewhat more lenient than pooling since "a miss is as good as a mile." With the sampling procedures the compensation is effected at the indicator or sub-indicator level. Therefore, it does not matter how close the teacher came to

receiving credit for the sub-category. In the pooling systems weaknesses in one set of data must be balanced by strong ones in the other data set. Thus, with pooling, the initial performance might be so weak that it cannot be balanced by a second. The choice between these two strategies will depend on the assumptions that are made about evenness of performance.

Summary

Obviously, there are many more aspects of multiple data sets. If some of the sub-categories are considered essential the rules change considerably. Also, if there are three observers or three occasions or both, the data set is more complex. However, working through the examples presented thus far should provide sufficient background to extrapolate into more complex situations. The purposes of the assessment should certainly be considered when the decision-making structure is planned. All of these systems allow some unevenness of performance. The way this is done and the uniqueness of the sub-categories are different.

Discussion

Most of the principles that have been discussed and illustrated are common sense ideas that have been developed based on a good deal of thought about the definition of minimally acceptable job performances. That is, all of the scoring systems are ways of describing how much a teacher must do in order to be scored positively--certified, continued in employment, or whatever. However, developers of performance assessment systems would do well to consider how many criteria, categories, sub-categories and the like a teacher can fail to demonstrate and still be called acceptable. The question might be phrased something like this "would you want to employ, certify or endorse a teacher who could not or had not done the following things adequately?" This additional

common sense concern might help determine if the system is too liberal since most of the emphasis has been on setting meaningful standards.

Finally, there is room for some "science" as well as a lot of common sense in this process. Since most of the criteria have been expressed as three out of four, we would do well to consider the meaning of the remaining one-fourth. Most of the previous discussion has treated this area as a common sense safety or slip factor for the teacher's benefit. An alternative, and perhaps more scientific, view would be to call this a confidence interval. In fact, the magnitude of this portion of the scores might be set as a confidence interval based on the reliability and standard deviation of the measures. Georgia has used this strategy in its system, considering that all of the elements are necessary and desirable. In such cases the confidence interval would not need to be set at an arbitrary or common sense level such as three fourths. Instead, it could be set at the level which would provide a tolerable level of possible false denials. If this were done with total scores, sampling teaching behaviors in our example, a relatively small number of teaching behaviors would be included in the confidence interval.

The following principles have been illustrated:

1. Items in relatively small categories are more potent than items in large categories in affecting a negative decision.
2. Teachers can "get away with fewer deficiences in a system with small categories."
3. A small confidence interval in a big category will still uphold standards.
4. Sampling across sub-categories causes sub-categories to lose identity.

5. Replication is a rigid procedure which can maintain even performance.
6. Pooling and sampling small elements also assure even performances. Sampling allows larger variations in some categories; pooling allows small variation in all subcategories.
7. Requiring particular elements--specific behaviors or higher order categories--greatly affects their impact on decision-making.

The reliability models that are associated with many of these procedures are complex. Yet, it is possible to determine the adequacy of measures for making the kinds of decisions needed. In fact, generalizability procedures can help determine the effect of alternative procedures (Capie, 1981) such as varying numbers of raters, occasions, items, etc. Additionally, the validity and usefulness of the data for various purposes. Whole scores may be more defensible for certification but inappropriate for staff development.

References

Borman, Walter C. Behavioral Approaches to Evaluating Individuals' Work Performance. Paper presented at the fourth annual Johns Hopkins University Symposium on Educational Research, November 1982.

Capie, W. Sampling considerations in classroom research. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, 1981.

Cooper, W. H. Ubitqrutons Halo, Psychological Bulletin, 1980, 10:218-44.

Jacobs, R. Format Variation for Rating Scales: Much Ado about Nothing. Paper presented at the fourth annual Johns Hopkins University Symposium on Educational Research, November, 1982.

Jacobs, R., Kafry, D. & Zedeck, S. Expectation of Behaviorally Anchored Rating Scales. Personnel Psychology, 1980, 33:595-640.

Landy, F.J. & Farr, J.L. Performance Rating. Psychological Bulletin, 1980, 87:72-107.

Rosenshine, B.V. & Furst, N. The use of direct observation to study teaching, in R.M.W. Travers (ed.) Second Handbook of Resarch on Teaching, Chicago: Rand McNally, 1973.